CURE for VIRUS:
Data Analysis Pipeline Overview

Niv Drory, Ralf Köhler, Jan Snigula, &
Helena Relke

MPE

```
 * Output:
 *     -best fit total flux of LAE on the sky in ADU from model: modflux
 *     -degrees of freedom used to calculate chi^2: c
 * Return:
 *     -reduced chi^2 of best-fitting model
 */
float get_chi( vector< list<ResElem> >& Model_LAE, float fdata_fmod,
               vector< MArray<float,2> >& Data, vector< MArray<float,2> >& Data_err,
               MArray<float,1>& flux, float& modflux, int& c )
{
   int npix = 0;
   float chi2 = 0.0f;
   float modflux_ = 0.0f;

   // loop over all frames/dithers/shots
   for( unsigned int s = 0; s<Model_LAE.size(); ++s )
   {
      list<ResElem>& shot = Model_LAE[s];
      float shotflux = 0.0f;

      // loop over all resolution elements
      for( list<ResElem>::iterator p =shot.begin(); p != shot.end(); ++p )
      {
         ResElem& res = *p;

         // if there's a certain flux in a fiber we bother looking into it ...
         if( res.flux > g.flux_thresh )
         {
            for( vector<Pixel>::iterator i = res.p.begin(); i != res.p.end(); ++i )
            {
               Pixel& pix = *i;
               const float e = Data_err[s](pix.x, pix.y);
               if(e == 0)
                  continue;
               const float d = Data[s](pix.x, pix.y);
               ++npix;
               shotflux += get_normflux( d, pix.x, pix.y );
               modflux_ += fdata_fmod*pix.flux;
               chi2 += pow2(fdata_fmod*pix.flux - d)/(e*e);
```

Data Analysis Software

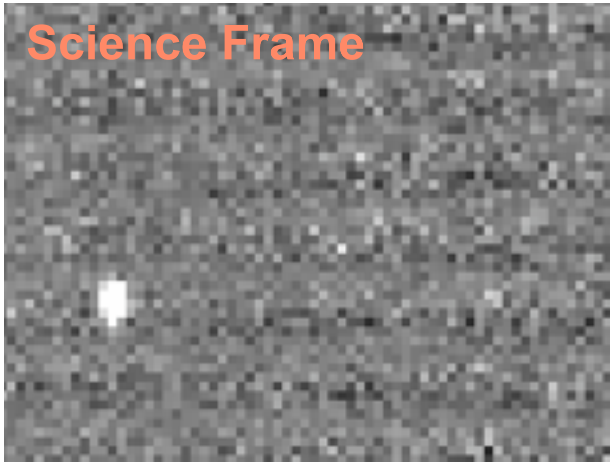
- **Quick-look reduction + visualization in real time (on site)**
  - astrometry, photometry of field stars, quality control and assurance
  - results available minutes after observation
  - monitoring, visualization, and feedback tools for observers
  - provide input for observation planning and scheduling during night
- **Full reduction within same day of observation (remote)**
  - basic calibration, registration, source detection, catalog creation, source classification
  - **emission-line object detection and extraction**
  - general source extraction
  - metadata tagging, database ingestion, public access
  - data quality monitoring: photometric standards, source counts, field-to-field variation, overlapping objects, etc. available online to be inspected
  - reduced data should be available within 8h

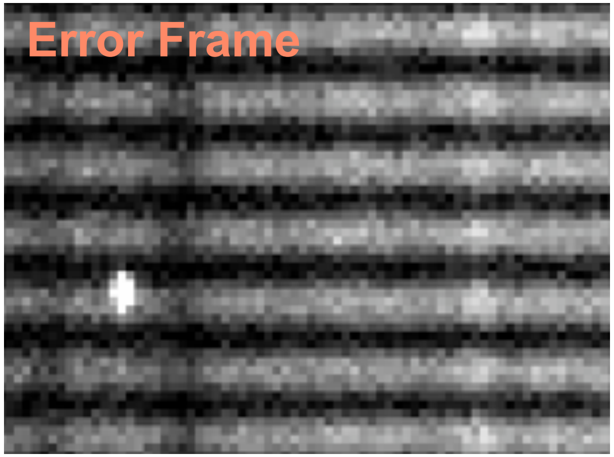
Data Analysis Principles I

Robustness of LAE detection crucial:

Each frame has an associated error frame which records the error in each pixel. This error is propagated through all computations.

- pipeline remembers flatfield, cosmics, sky noise, ...
- accurate per-pixel significance measures
- statistically accurate and robust *object detection significance*
- *stable and well-defined detection limits across survey*
- developed for and in use for microlensing (Goessl & Riffeser 2002)





Data Analysis Principles II

Automated LAE detection in 2D, non-resampled frames:

- preserve the original *data* and *noise characteristics:* crucial for accurate statistical modeling of the data - low S/N detections
- make use of full information available on the CCD: use 2D image information for source classification and extraction
- robust detection, well-defined significance measures possible
- images are never resampled
- median sky is extracted from data frames
- sky is resampled to geometry of each frame for subtraction

Astrophysical classification of continuum sources is better done on extracted 1D spectra: if necessary, resample only one at the end of processing.

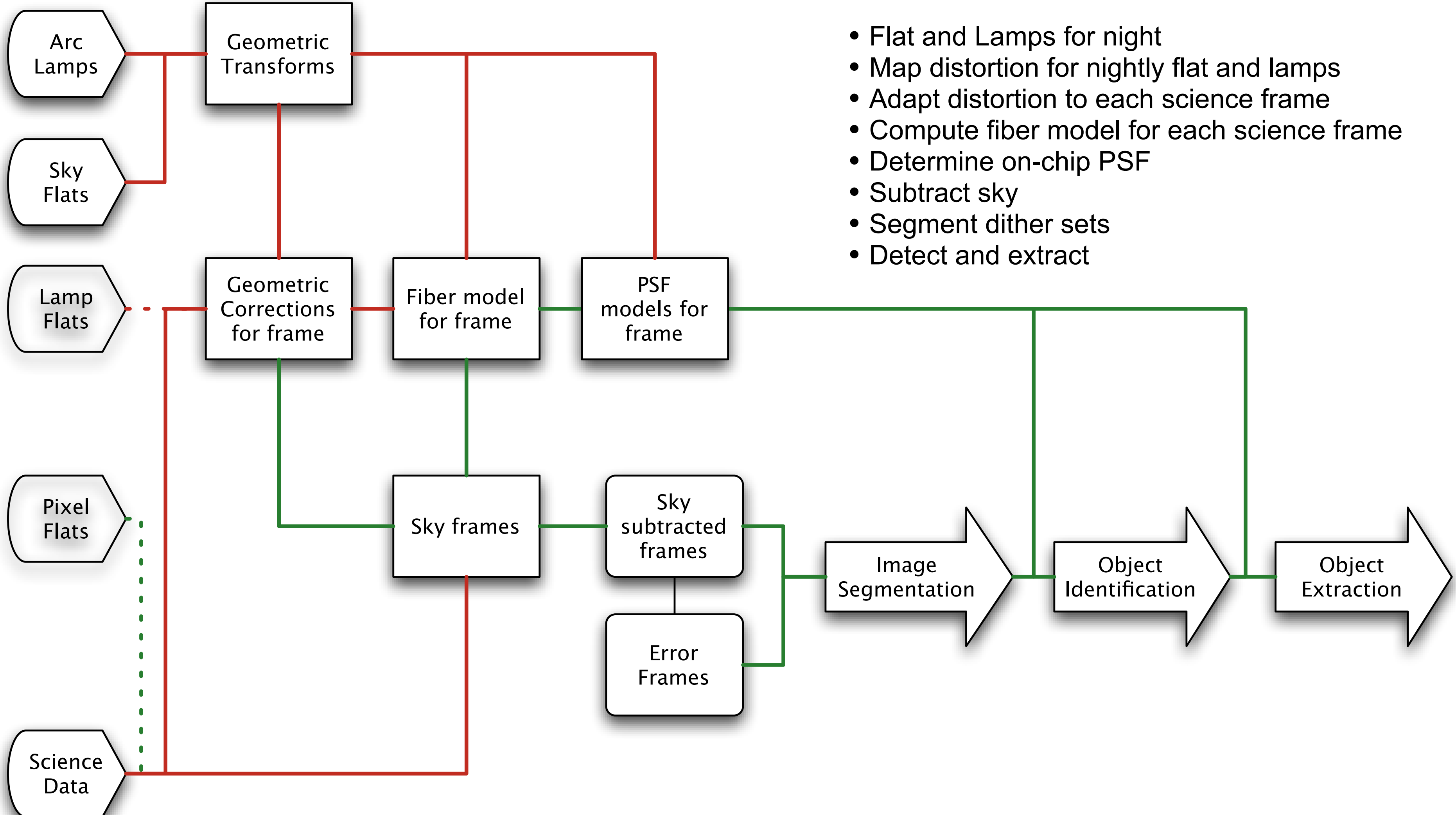
Data Analysis Principles II

We need to reduce data from 192 almost identical spectrographs:

- use as much knowledge about the spectrographs as possible
- distortion pattern, dispersion model, illumination, PSF variation, ...
- then parameterize pipeline and “fit” to each spectrograph
- monitor and optimize with each night’s calibration data
- provides in-depth long-term monitoring of instrumental changes
- identify systematic trends in the data

Pipeline Overview

- Flat and Lamps for night
- Map distortion for nightly flat and lamps
- Adapt distortion to each science frame
- Compute fiber model for each science frame
- Determine on-chip PSF
- Subtract sky
- Segment dither sets
- Detect and extract



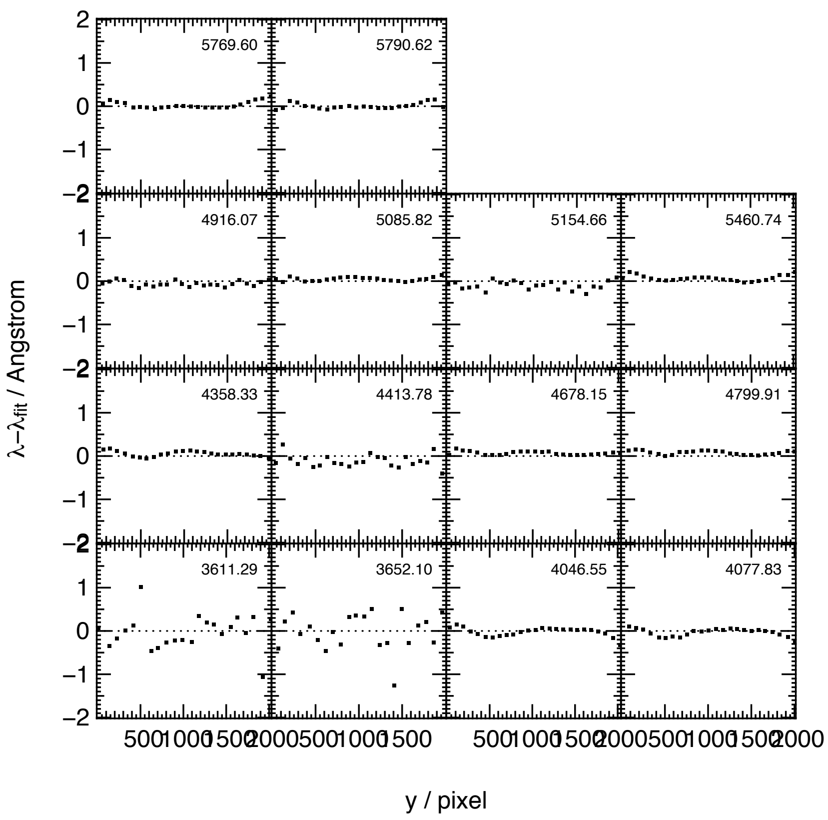
Wavelength Calibration

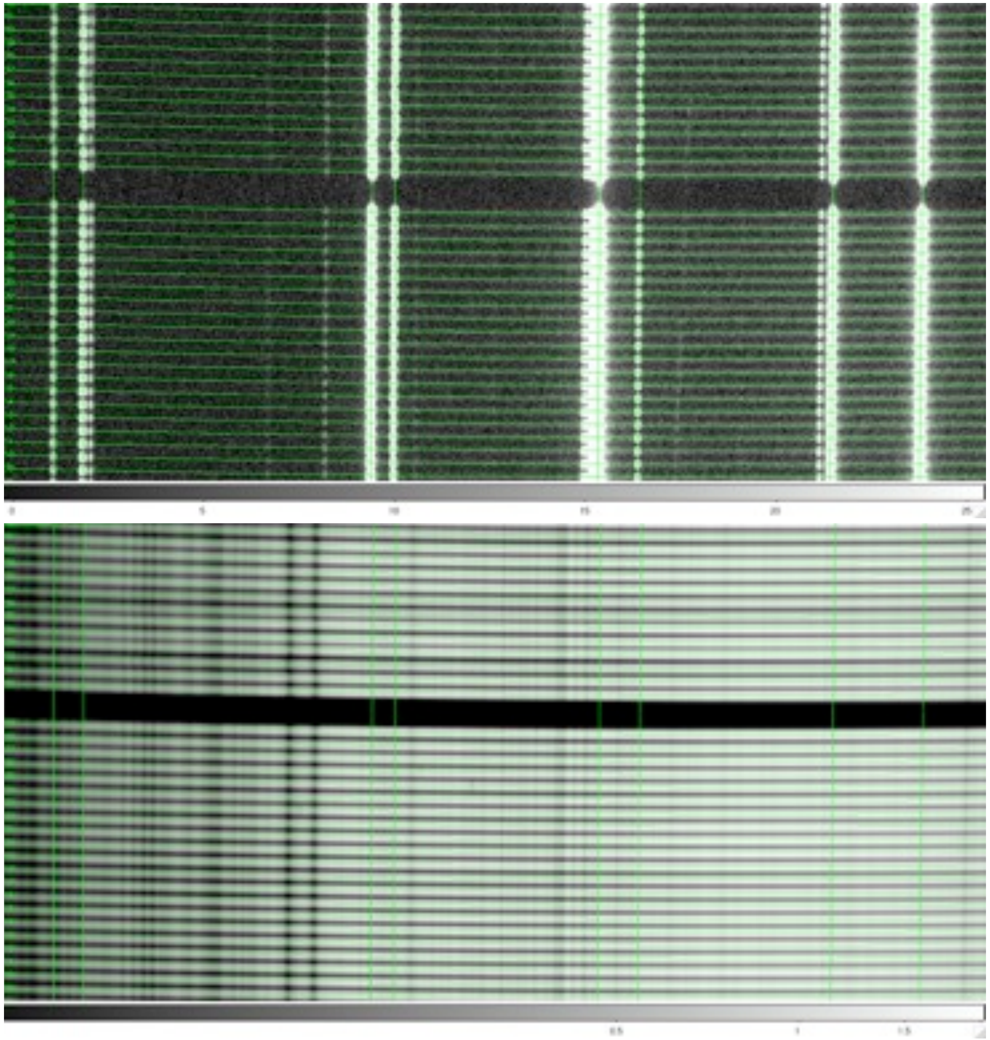
Wavelength solution, geometric distortion

- Fit global geometric distortion pattern; correct for fiber-slit imperfections
- Create a unique mapping between CCD {x,y} - {λ, f} - sky {α,δ}
- λ-coordinate: constant along lines of equal wavelength on CCD
- *f*-coordinate (fiber): constant along fibers

Wavelength and Geometric Calibration

Global accuracy ~ 0.1Å,
0.02 res elem.





Flatfields & Weights

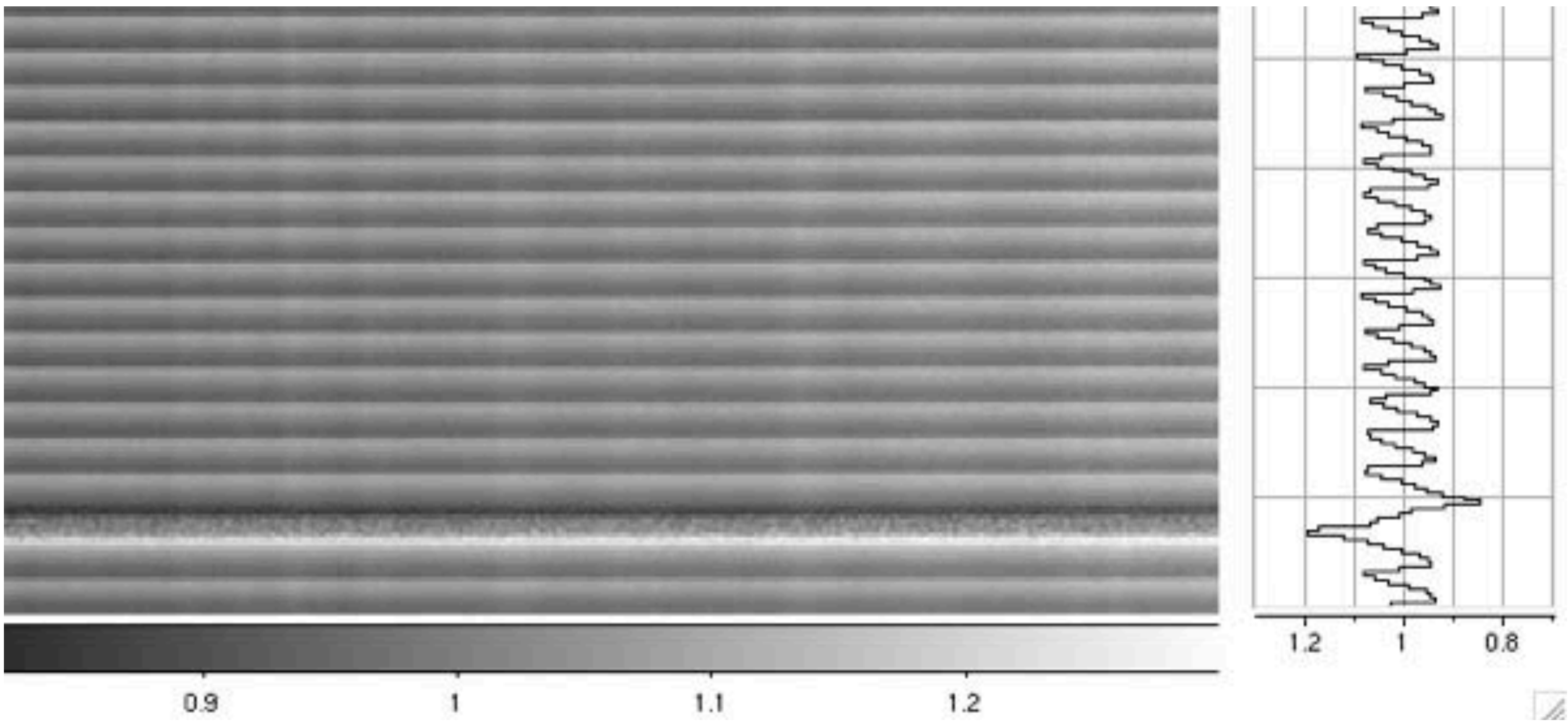
- Rather than divide by fiber-profile, multiply-in fiber profile to models or use as weights in extraction
- Sometimes, we want the contribution to a pixel of only one fiber (for point-source extraction)
- Sometimes, we want the total flux arriving at a pixel: extended sources, sky
- So we need a full model of the fiber profiles as a function of {λ, f}

- A flatfield exposure only contains the combined flux from adjacent fiber profiles
- We need to fit and separate the fibers in a flat or use individually illuminated fibers to get at the profile information

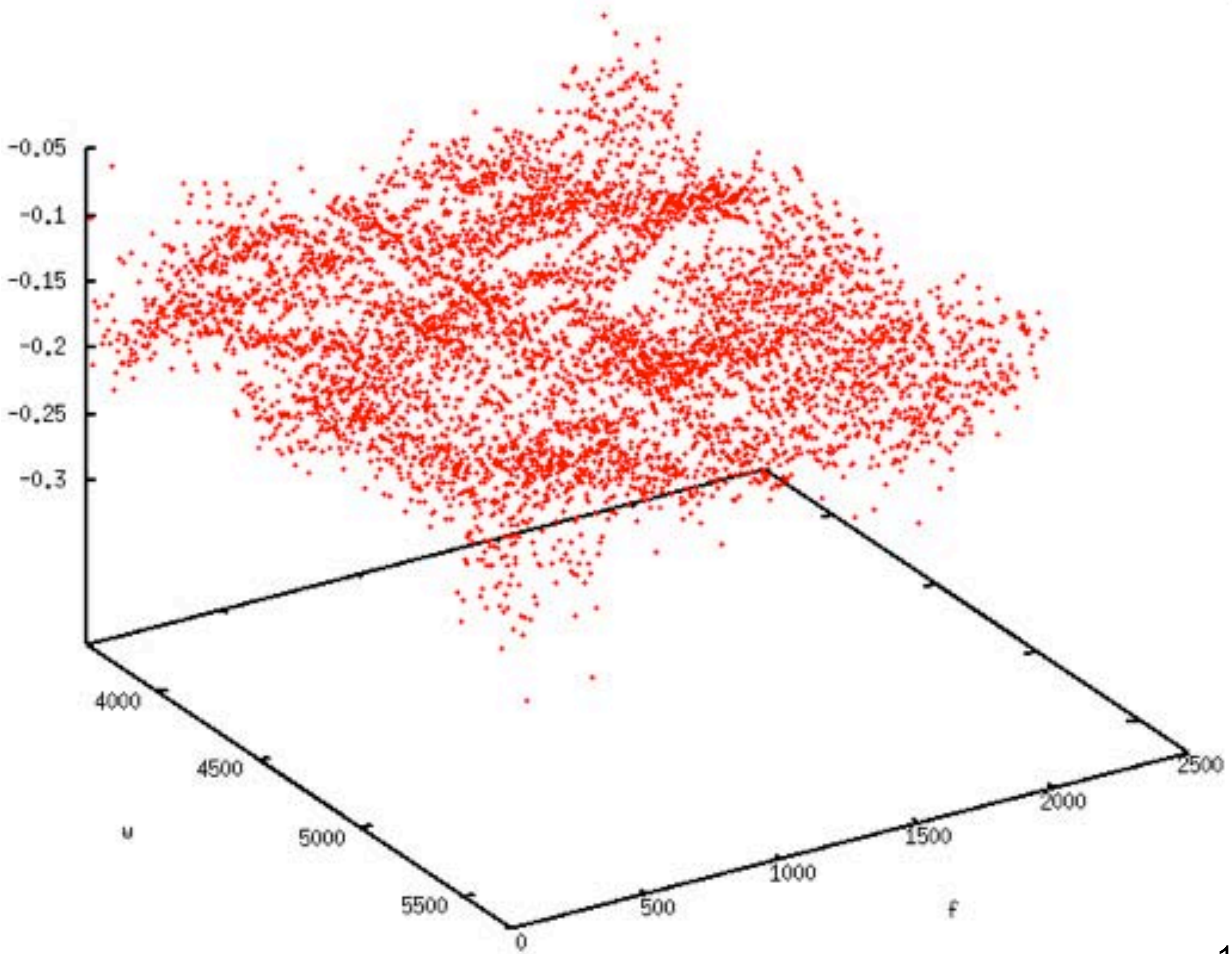
- Also, the fibers shift and so the flatfield changes ... flatfield cannot be used as extraction weights directly

Flatfield Variability I

- Compare evening and morning sky flats
- 10% difference in pixel weights
- Asymmetric pattern points to shift in fiber positions

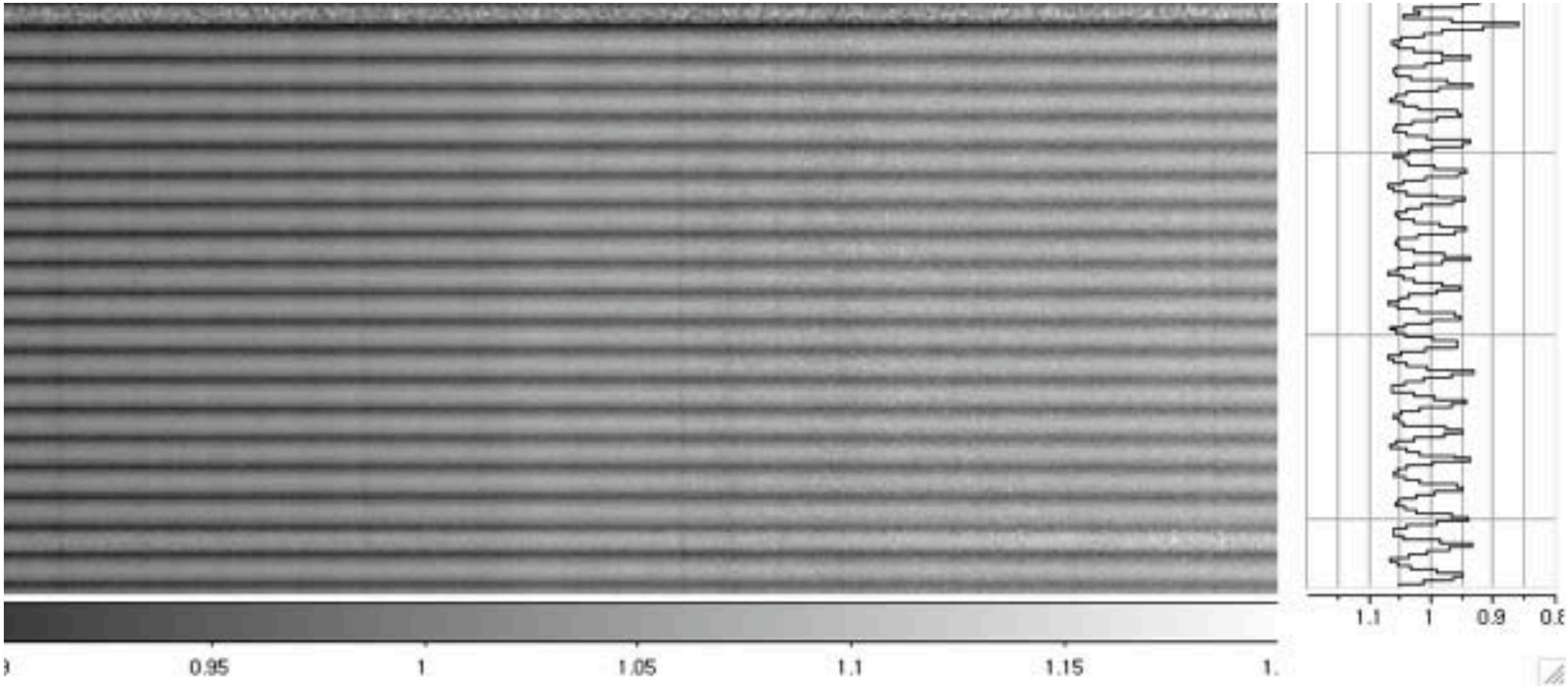


Flatfield Variability II

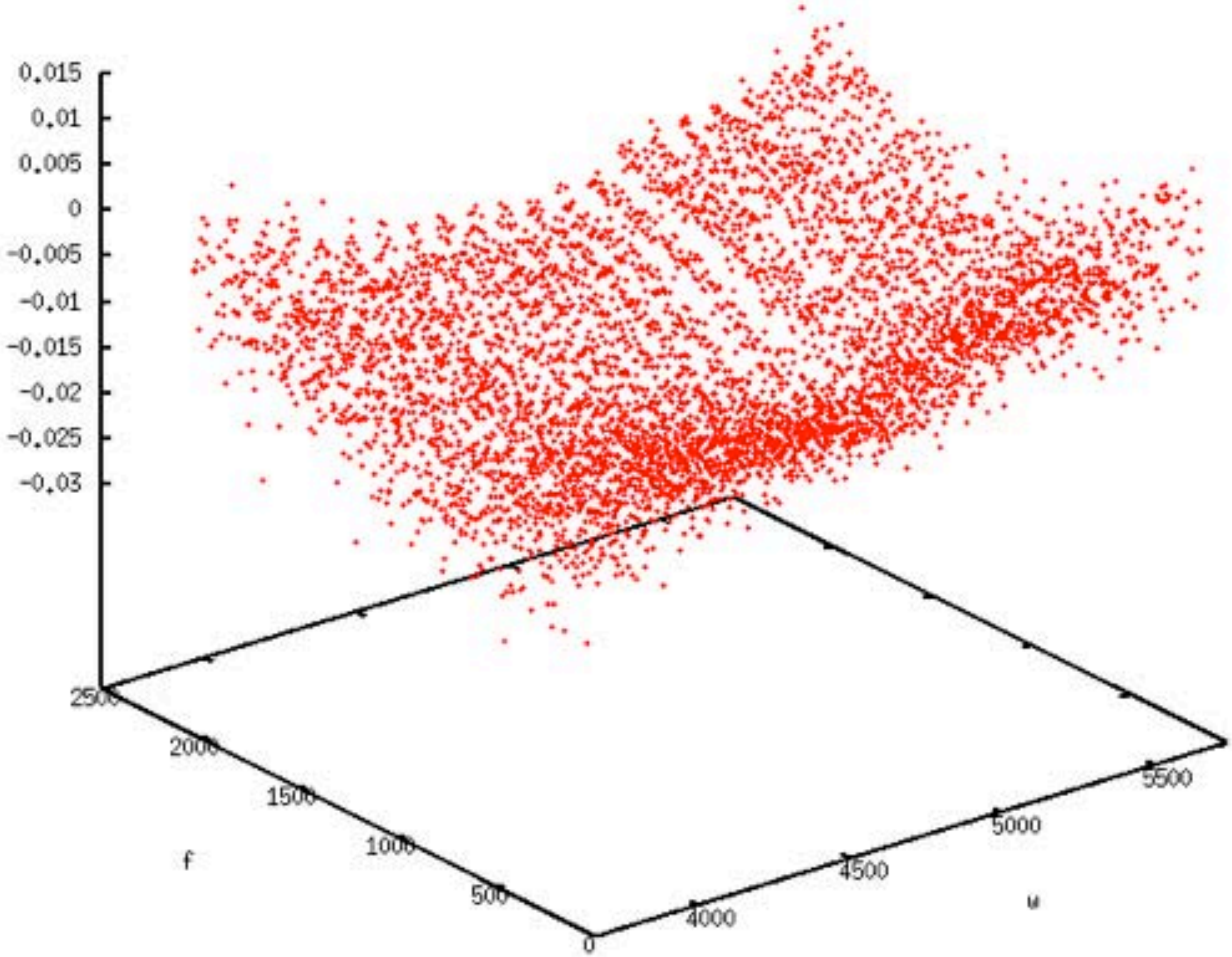


Flatfield Variability III

- After correcting for change in distortion pattern (shifting fibers to of evening flat to positions of fibers in morning flat using fiber model)
- Down from ±10% to ~ ±5% errors in weights
- Symmetric pattern points to difference in image FWHM!

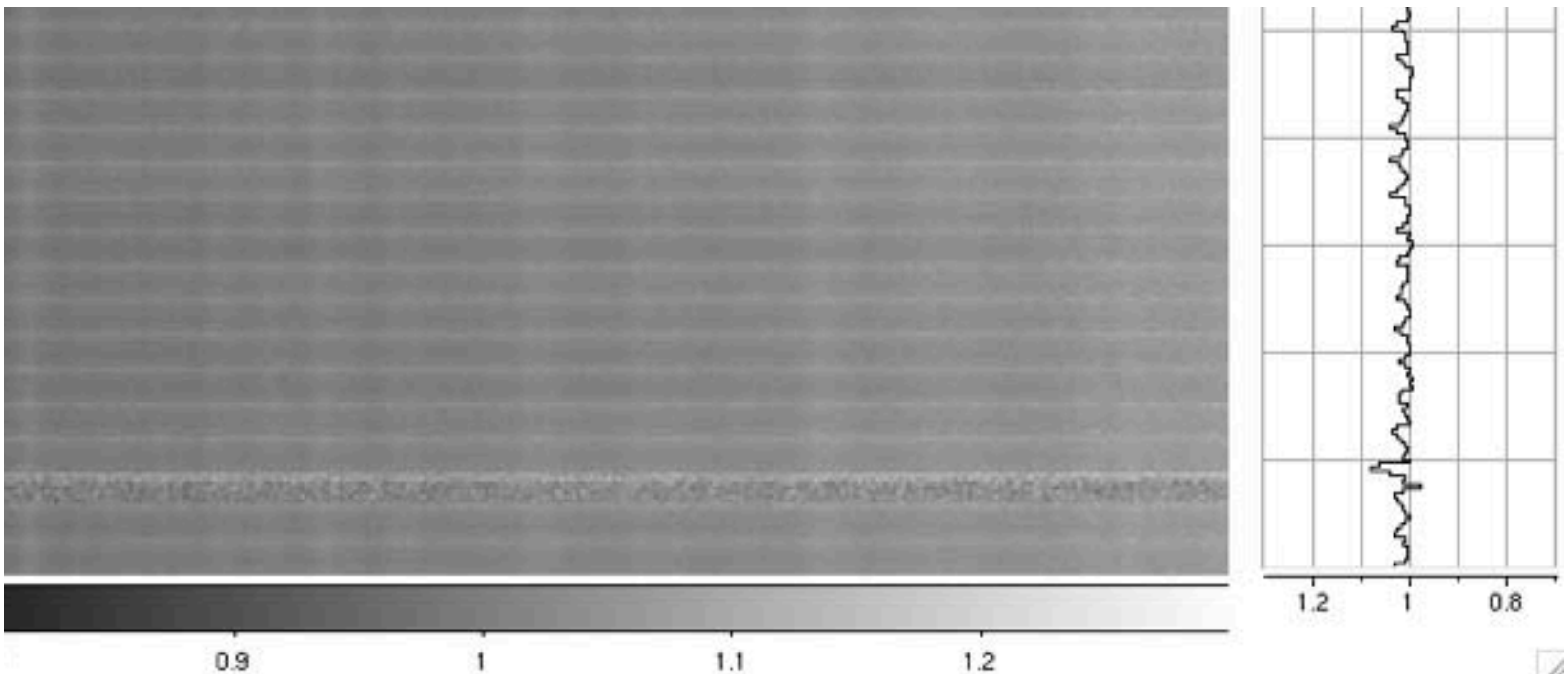


Flatfield Variability IV



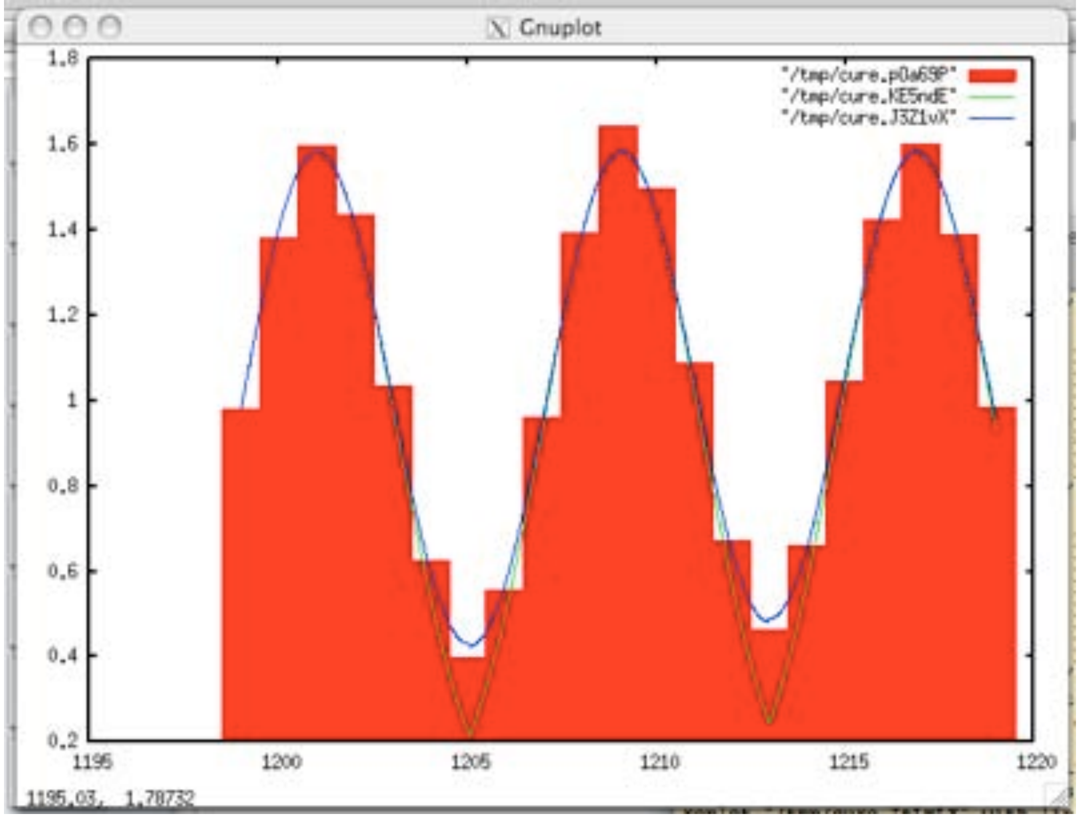
Flatfield Variability V

- After improving fiber models using individually illuminated fibers (stars)
- Down to ~ 2% rms errors in weights, except for extreme wings (5%)
- We can “recreate” an artificial flat for each science frame using the sky signal in the frame to adjust the geometry. NOT possible at HET!



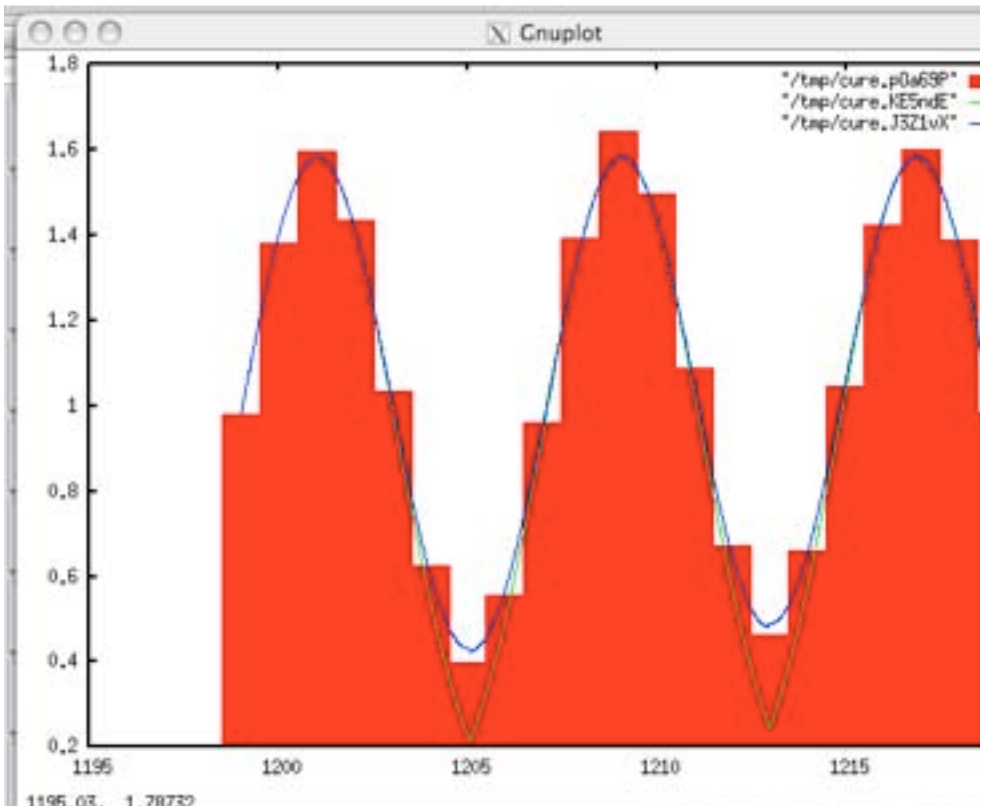
Fiber Profile Models I

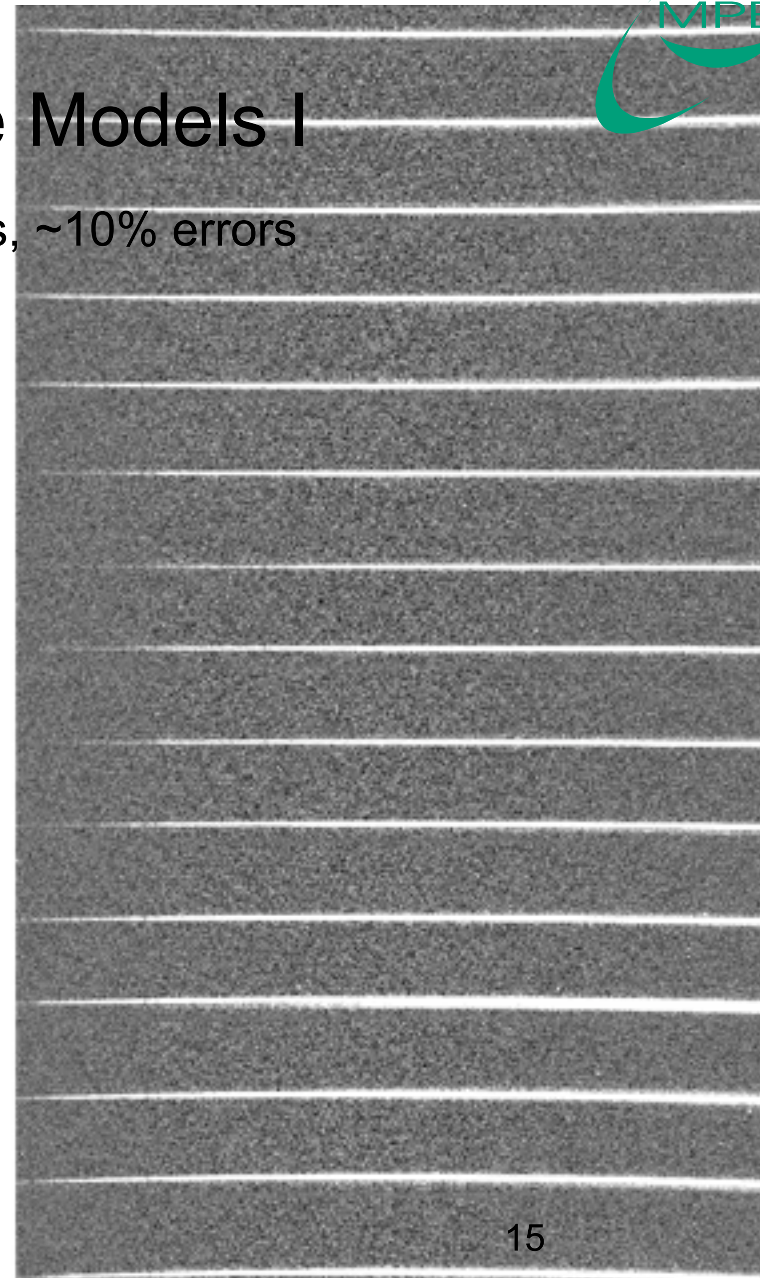
- Fiber profile models by fits to the sky flats, ~10% errors



Fiber Profile Models I

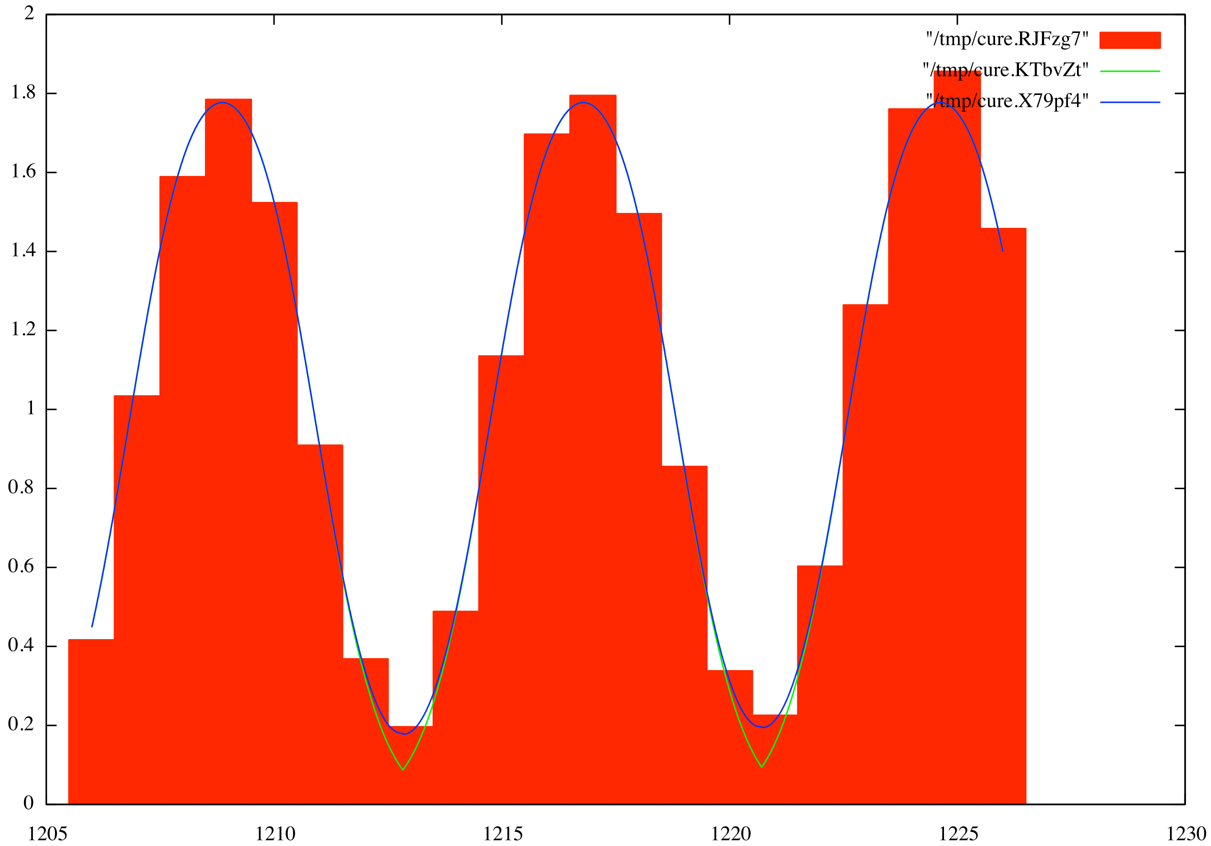
- Fiber profile models by fits to the sky flats, ~10% errors





Fiber Profile Models II

- Fiber profile models by fits to the sky flats compared to recalibrated fitting functions after evaluation of individually illuminated fiber images
- ~ 2% errors, further work on the fitting function required.



Sky Subtraction I

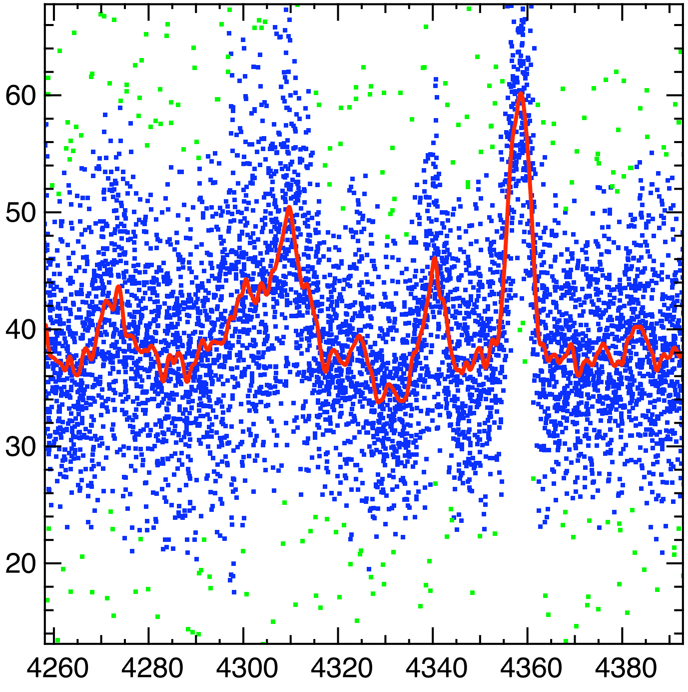
Sky subtraction

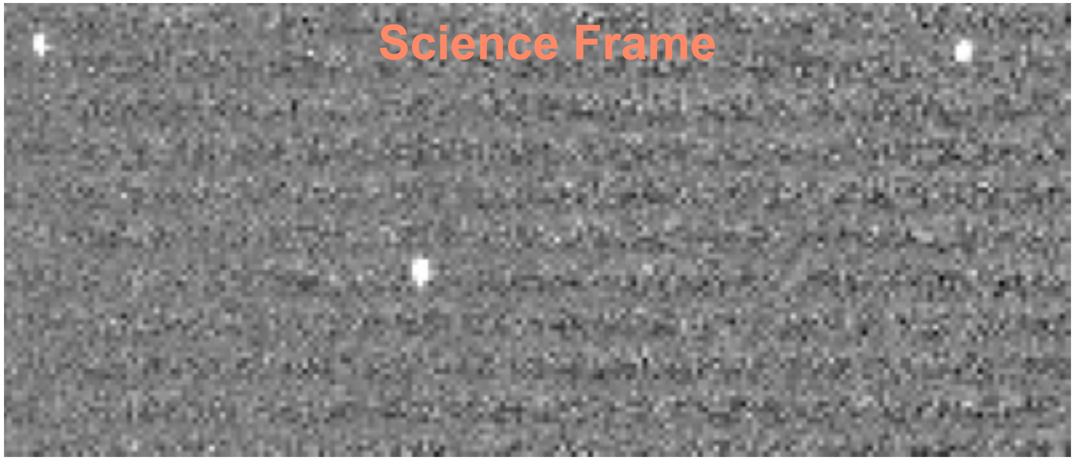
- Subtract “drizzled” high-resolution sky spectrum generated by combining information from pixels from many fibers
- Sort pixels by wavelength
- Fit approximating spline to pixel data
- Spline contains information at higher resolution than any single sky spectrum
- Evaluate spline at exact wavelenth of each target pixels to subtract sky using weight from fiber model
- The geometric distortion of the optics is, in fact, of great advantage here

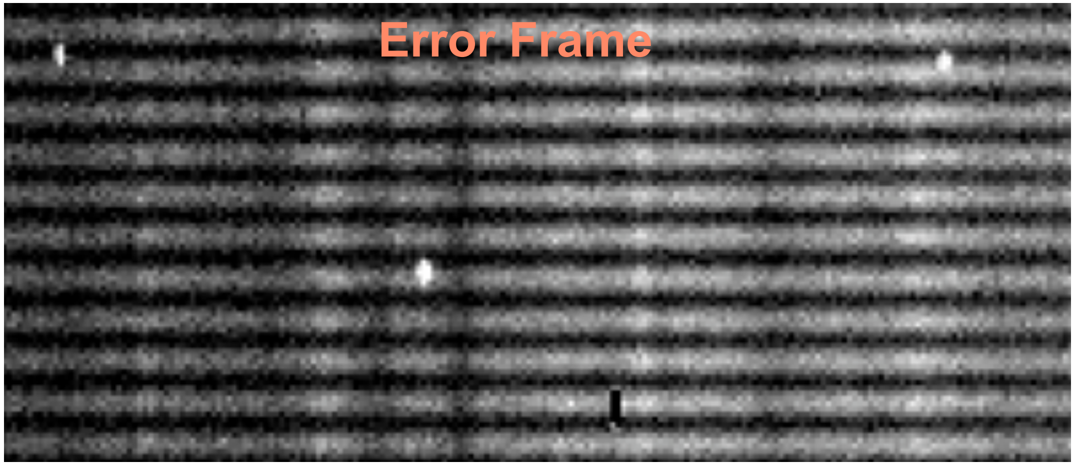
Sky Subtraction II

Sky subtraction by sorting individual pixels according to their wavelength and fitting by an approximating spline.

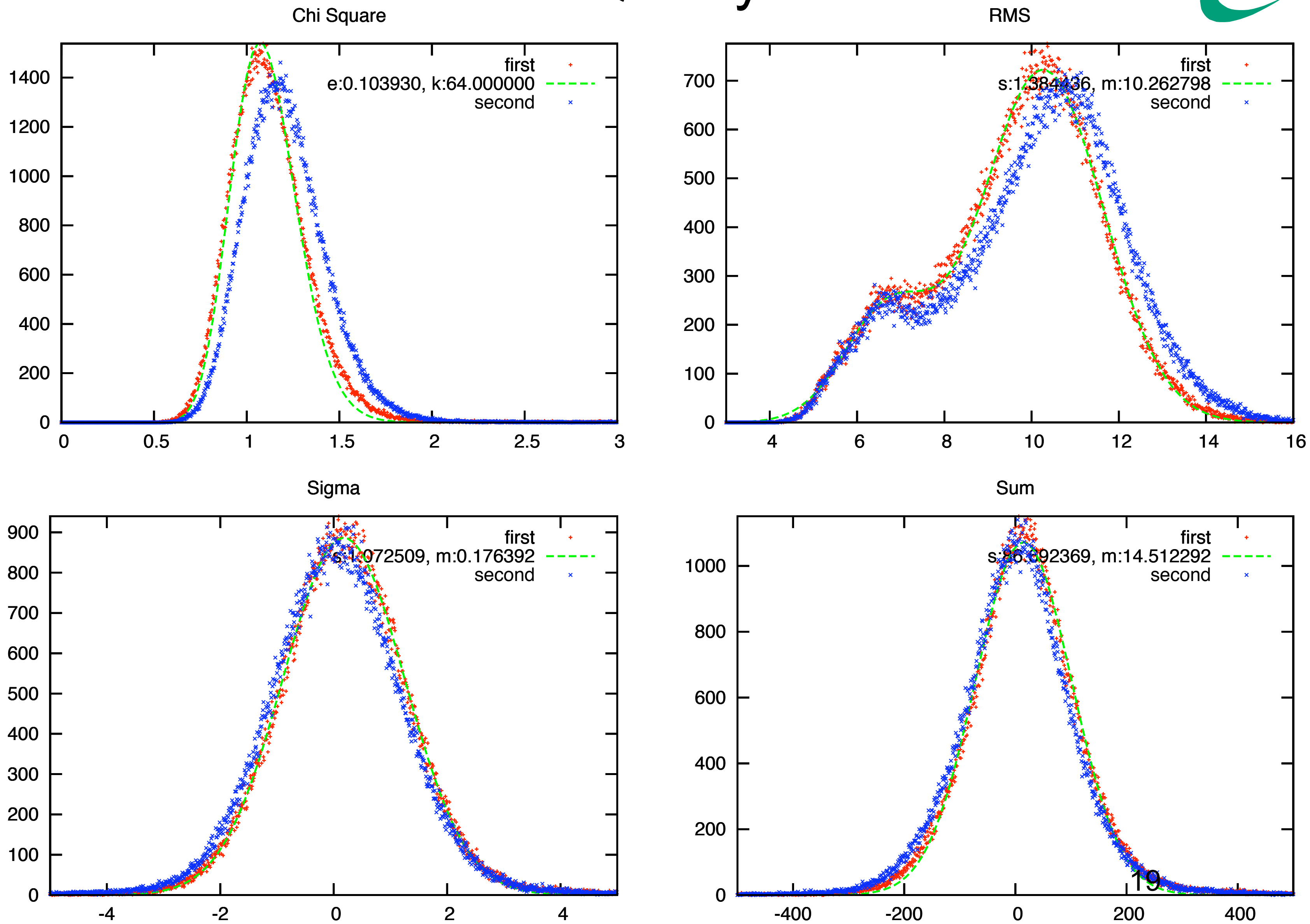
Sampled Sky + Spline







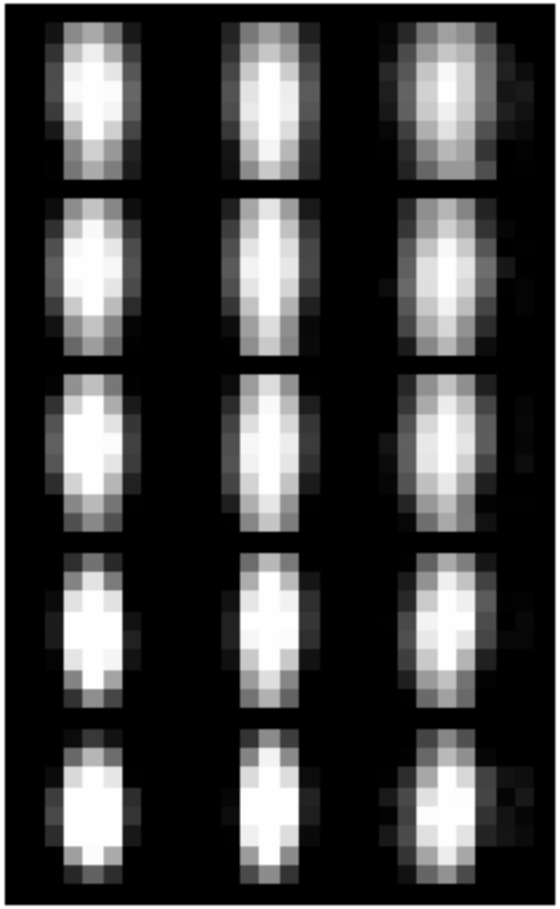
Final Quality Control



PSF Extraction

Spectral PSF Extraction

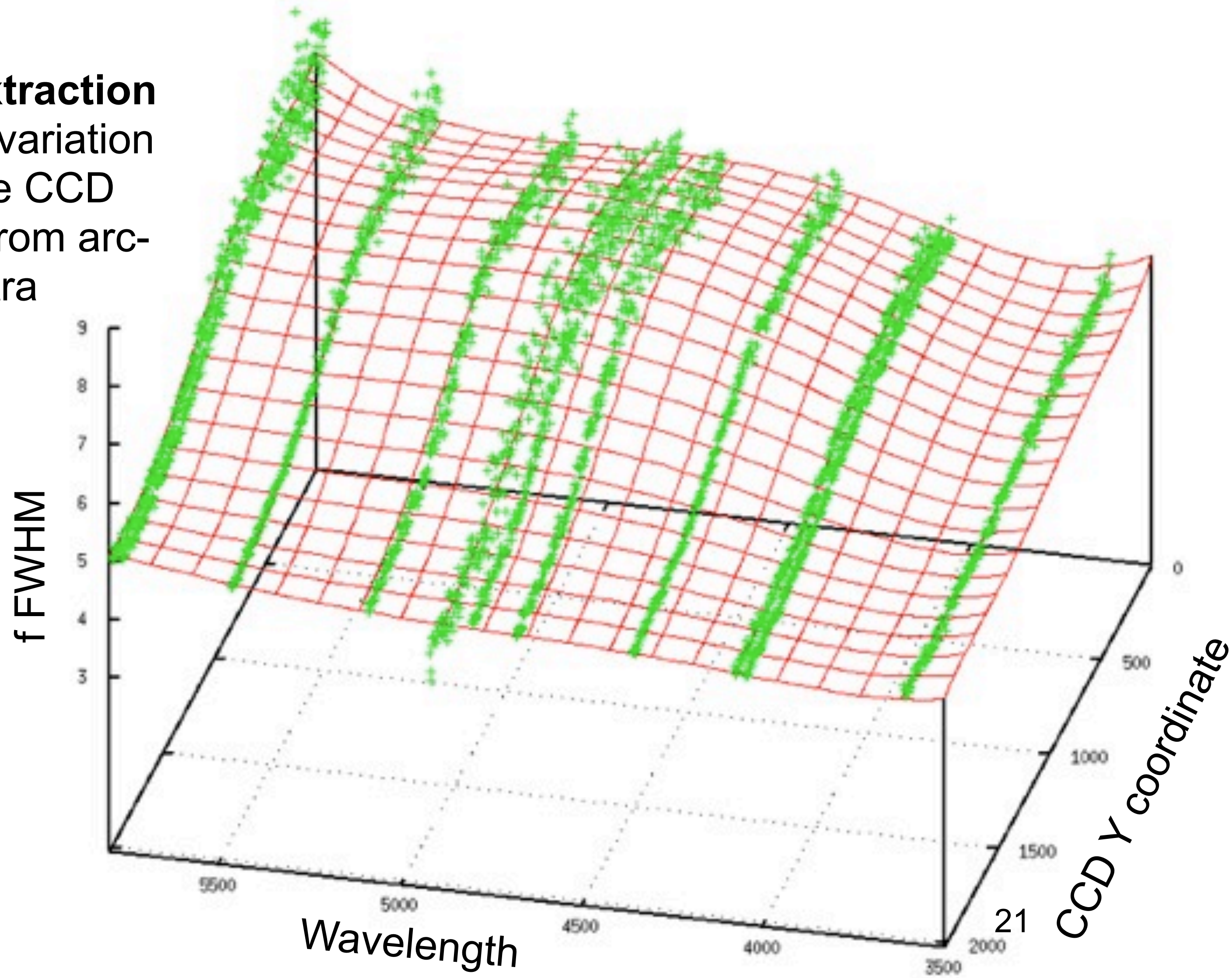
- PSF model and variation extracted from arc-lamp spectra
- PSF as a function of ccd {x,y} is generated from a grid of images of point sources extracted from arc-lamp frames.
- These can be either used directly (interpolating the nearest samples to the point of interest) or modeled as a function of some continuos functions CCD {x,y}.



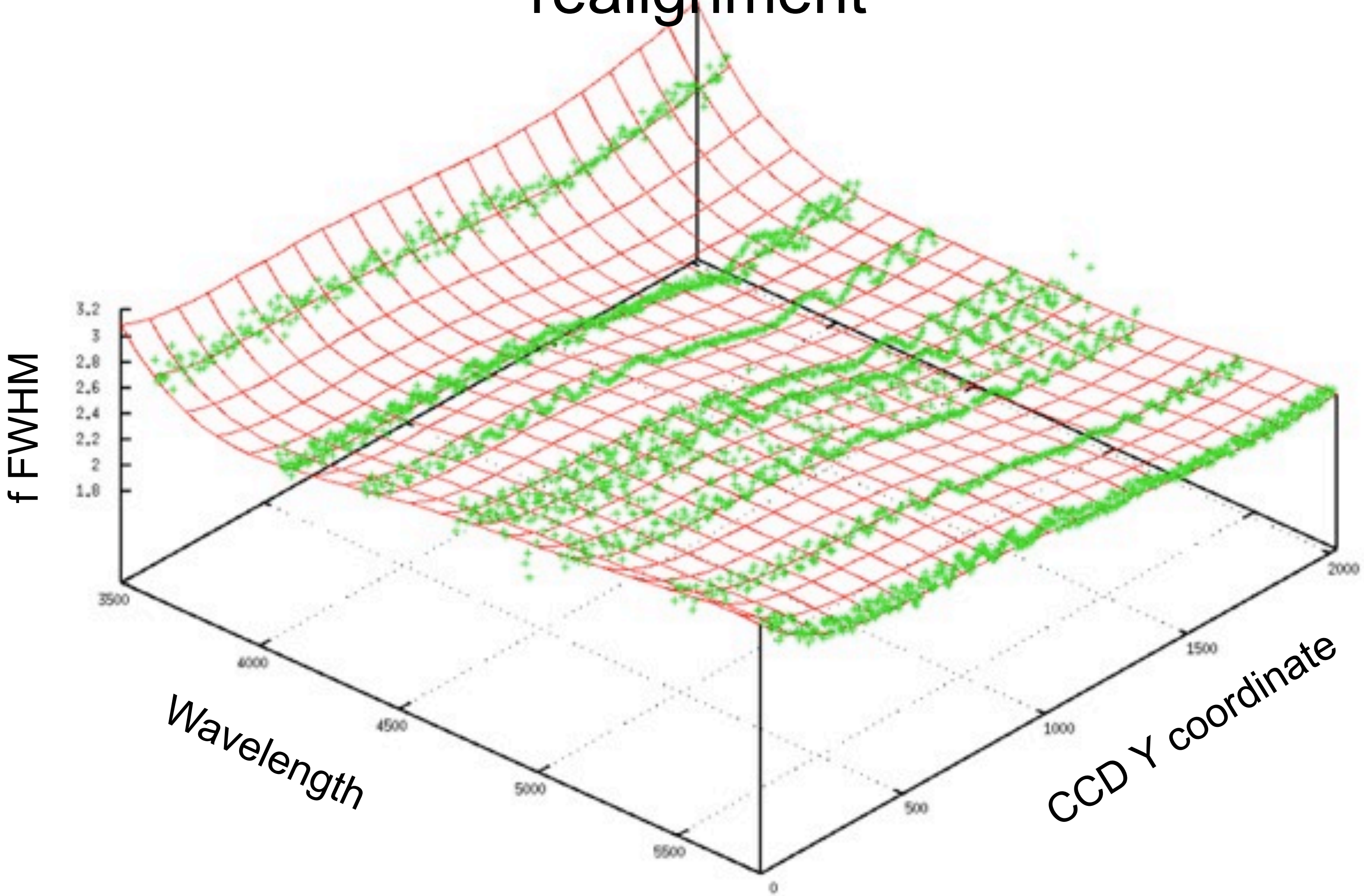
PSF Extraction

Spectral PSF Extraction

- Fit to PSF variation accross the CCD extracted from arc-lamp spectra



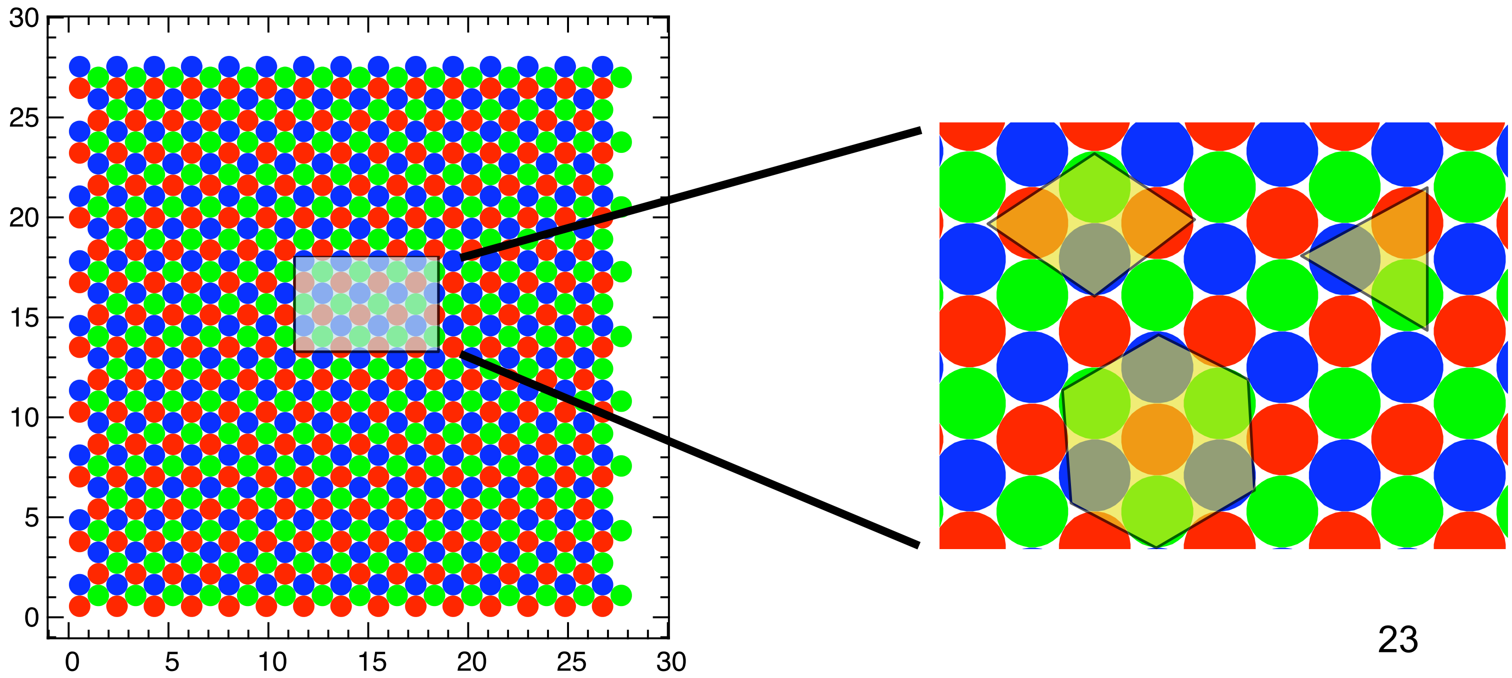
PSF Extraction - post realignment



Main Focus on Object Detection

We do no know beforehand where the objects are!

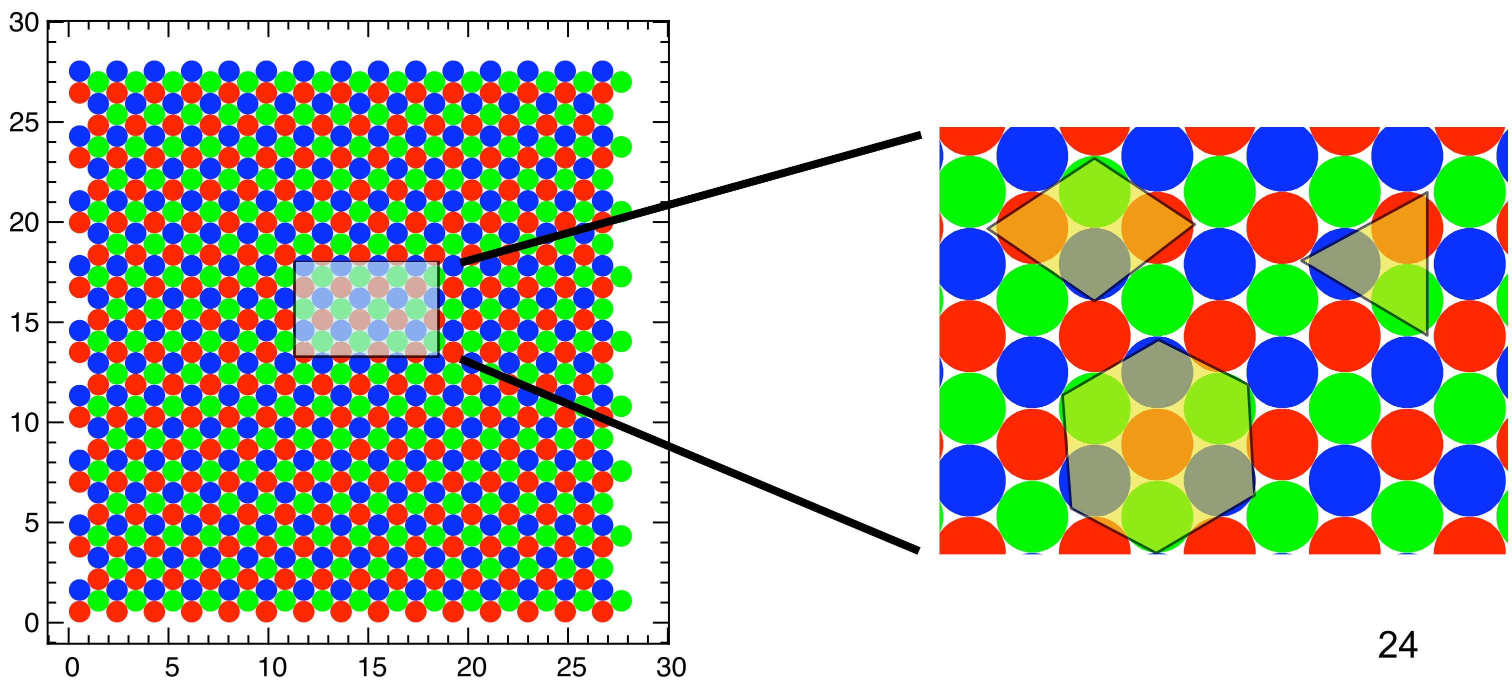
- HET PSF ~ fiber size: objects appear in multiple fibers!
- fit local maxima in the spectra? some objects not detected in individual fibers, only in combined spectra
- but which fibers to combine (and what are the weights)? more than one possibility depending on the object's assumed position!



Main Focus on Object Detection

Bayesian approach:

- What is the likelihood for a source at {x,y,z} on the sky given the data?
- Can be estimated for all {x,y,z} given a good noise model for the data.
- Robustness of LAE detection is crucial for DEX science.

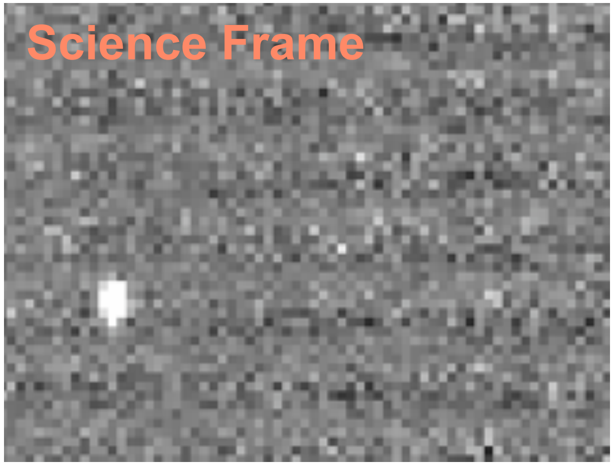


Main Focus on Object Detection

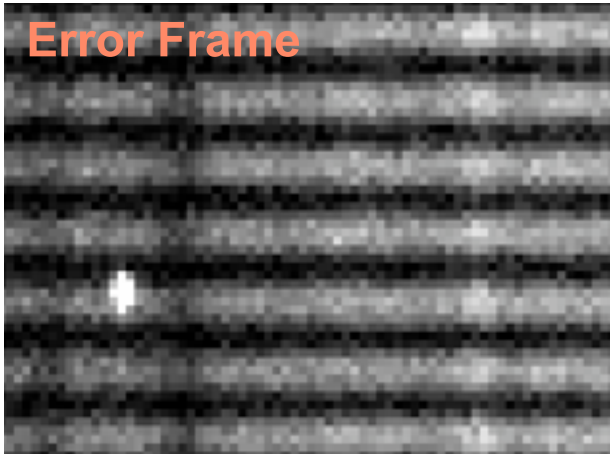
Difference to object detection in imaging:

- Typical imaging detection algorithms can assume background noise is homogenous on scale of small objects.
- This is NOT the case for spectroscopy: the sky has many high-frequency features comparable to object size (spectr. resolution).

Sky-subtracted spectra



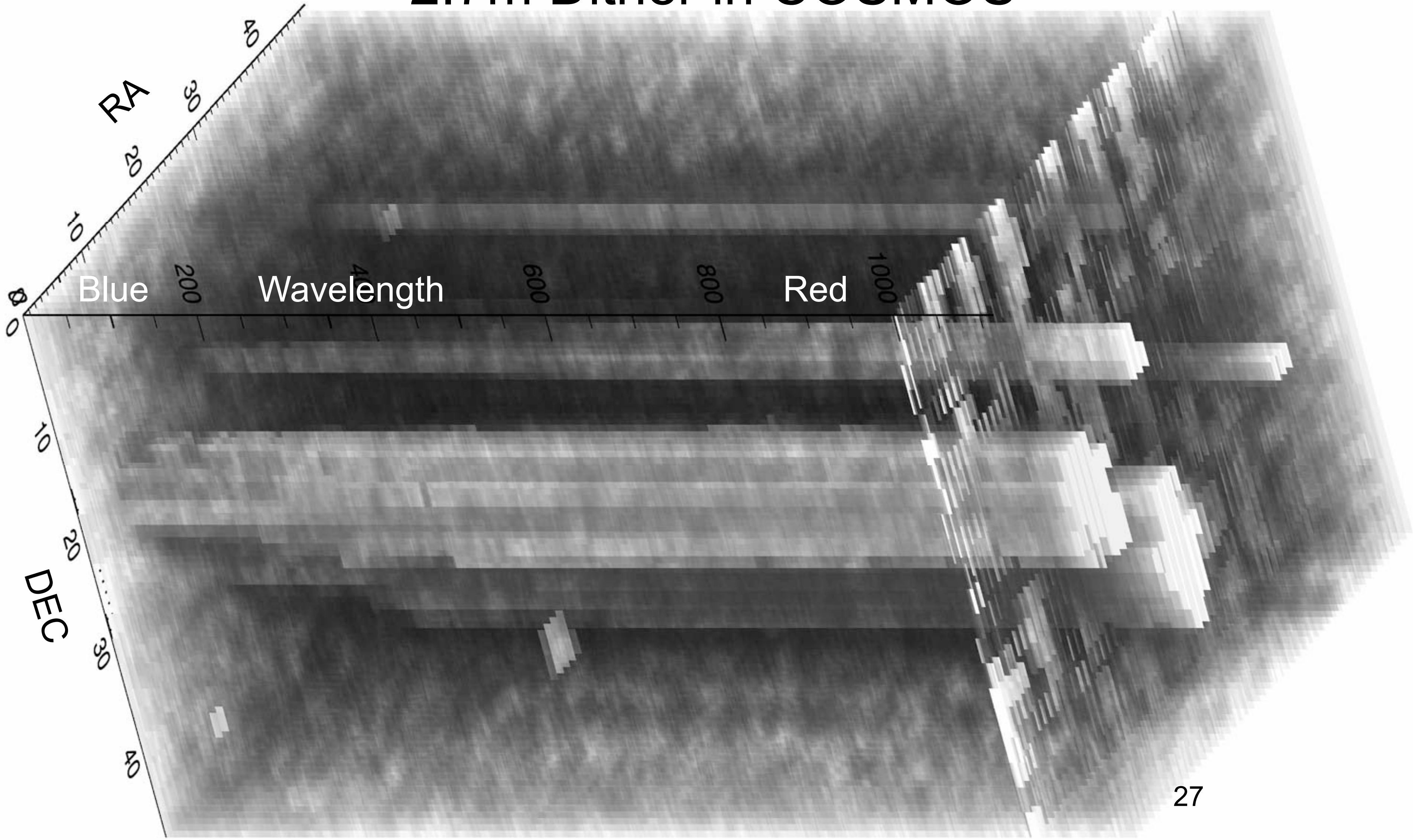
Photon noise



Source Detection and Extraction

- We detect on dither sets (multiple spatially adjacent frames filling the sky with fill factor ≥1; 3 at the HET, 6 on the 2.7m)
- Different sources (galaxies, stars, emission lines) will have different morphologies in (x,y,lambda)
- Segment the image: find consecutive resolution elements with significant flux in a dither set
- Look at morphology in the segmentation map to decide which extraction and classification algorithm to use for each source
- Currently only PSF extractor implemented
- Provide powerful interface for non-DEX optimized extractors: the `SegmentList`

The Segmentation Map: 6 Frame 2.7m Dither in COSMOS



Creating the Segmentation Map

1. Assume {α, δ, λ} on sky

2. Select all resolution elements in all CCD frames affected by this source:

   Given dither pattern, telescope PSF, and seeing, compute the flux fractions in each affected resolution element.

3. Compute significance above null-hypothesis using all the flux in these resolution elements - initial σ guess (ignores flux distribution for the moment) and write σ to segmentation map.

4. Segment the map and create Segment List.

(Point-) Source Detection

5. Walk the segment list: determine segment “shape” parameters.

6. if a segment has geometry consistent with a point-source and if σ above some threshold:

   Create full pixel-level model of a point source around the segment centroid and compute χ^2 against pixel data and errors.

   Record best-fit source flux, position, and likelihood.

7. Repeat with all segments and source types.

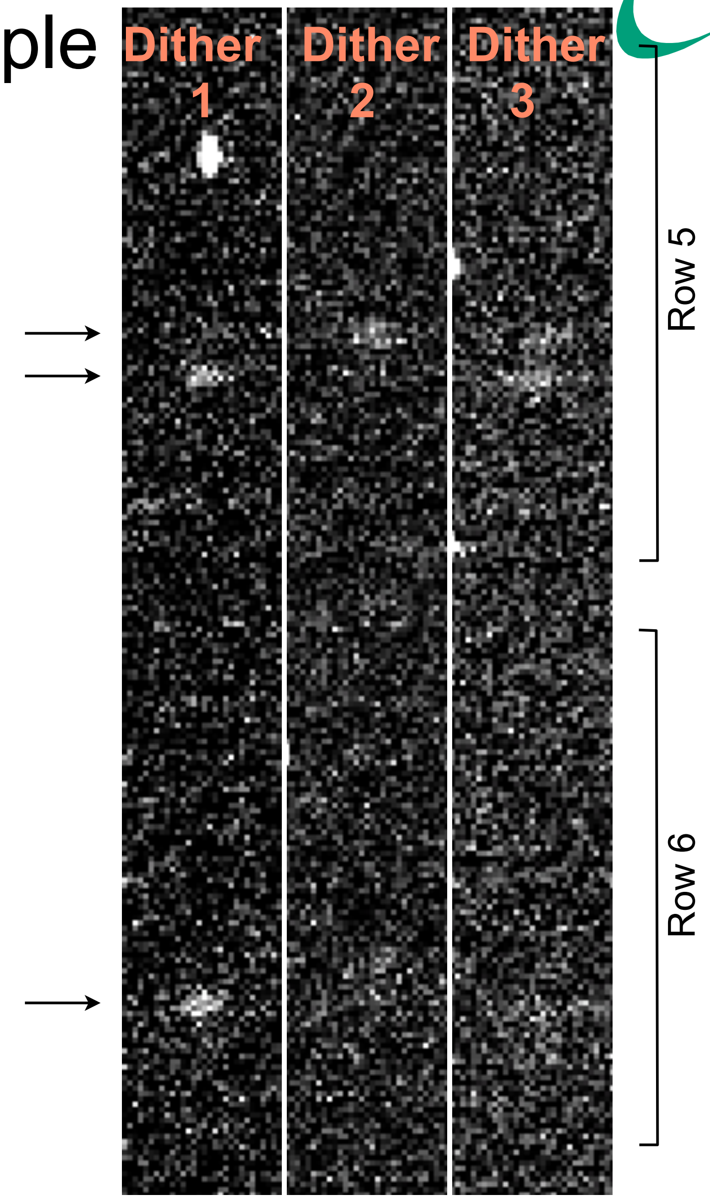
Remember: Seeing FWHM ≈ fiber diameter. This is not a “single point” detection. The complex pattern projected onto the CCD is very rich in information!

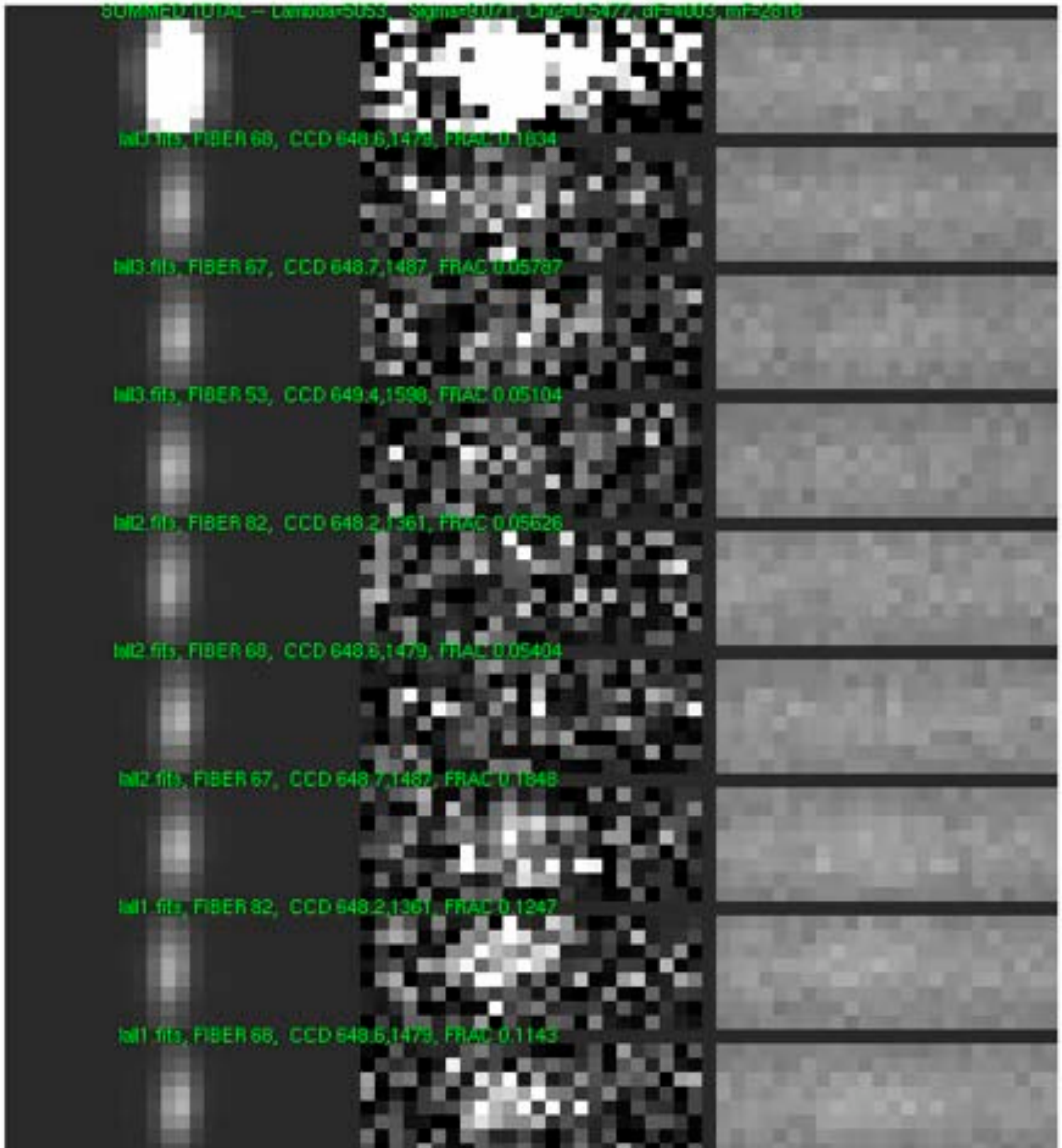
Object Detection - Example

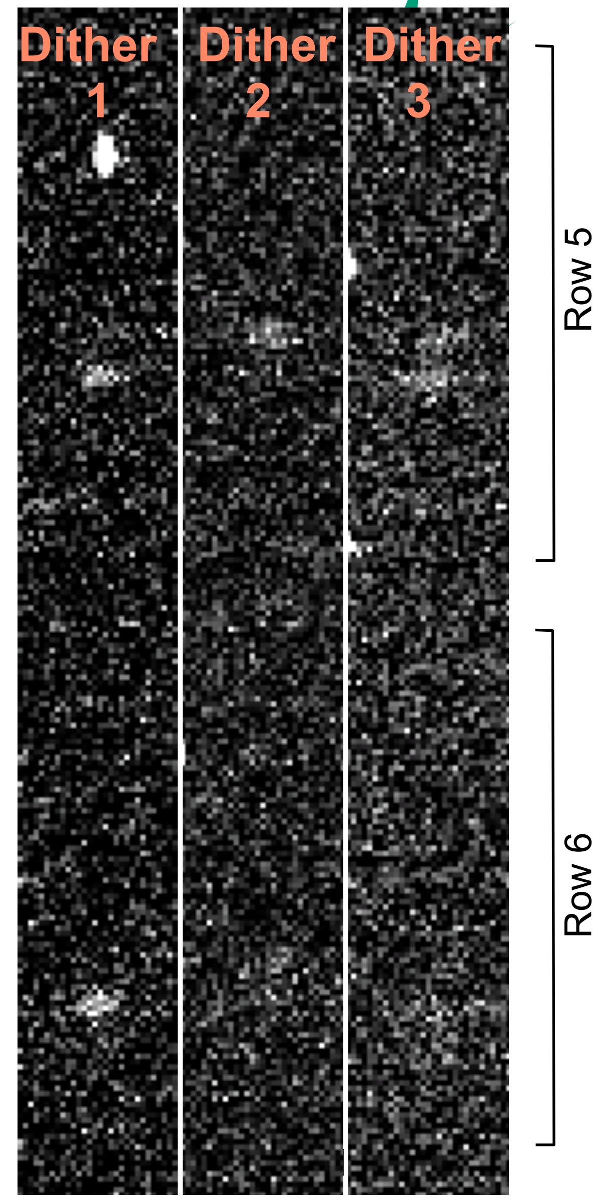
- Faint LAE detected in VP-HET data

- neighboring fibers: adjacent RA on sky
- neighboring fiber rows: adjacent DEC

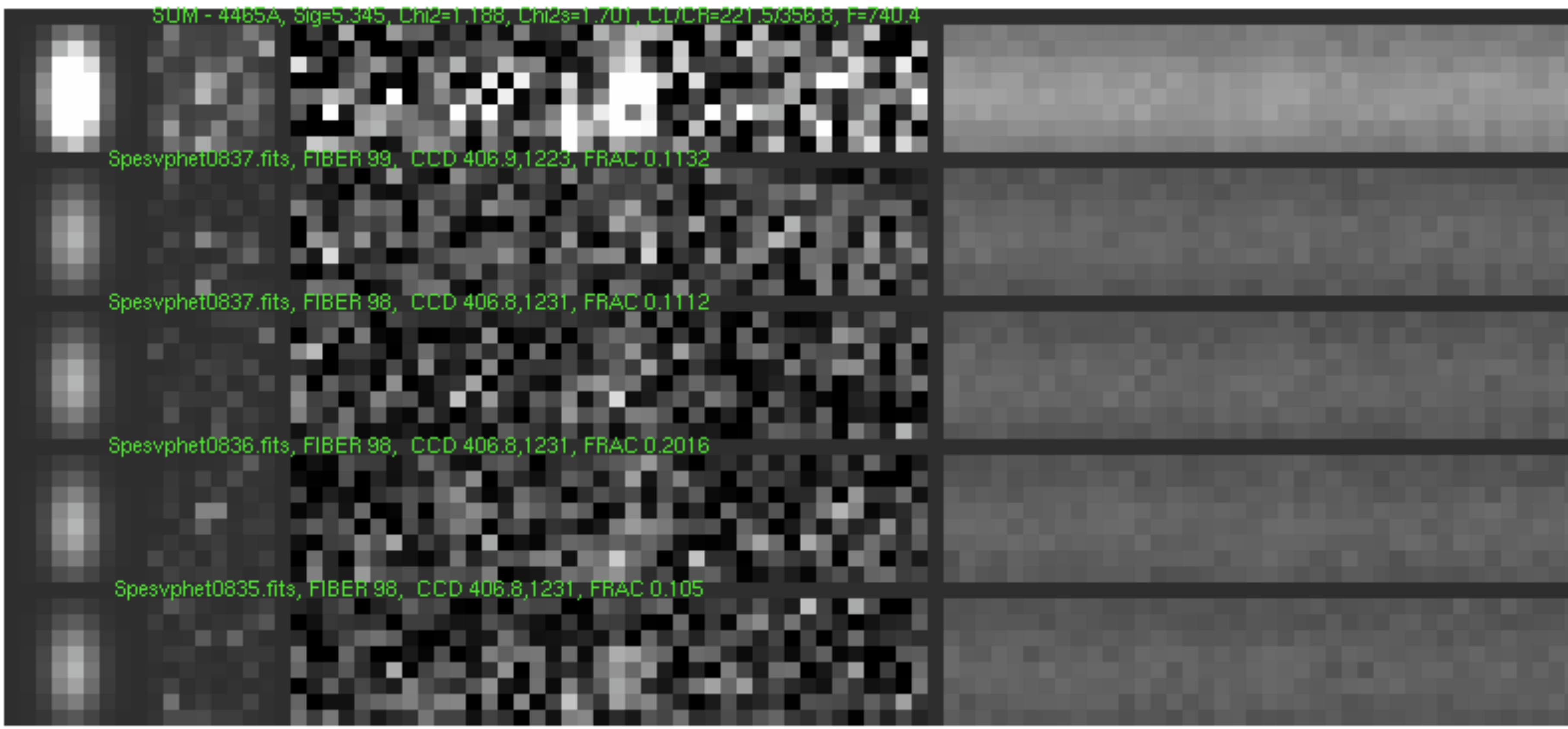
- Flux ratios between fibers pin-down position
- Flux distribution with resolution elemnts and between fibers discriminate between line sources and cosmic rays




SUMMED TOTAL
FIBER 68, CCD 648.6,1479
FIBER 67, CCD 648.7,1487, FRAC 0.05787
FIBER 53, CCD 649.4,1598, FRAC 0.05104
FIBER 82, CCD 648.2,1361, FRAC 0.05626
FIBER 68, CCD 648.6,1479, FRAC 0.05404
FIBER 67, CCD 648.7,1487, FRAC 0.1848
FIBER 82, CCD 648.2,1361, FRAC 0.1247
FIBER 68, CCD 648.6,1479, FRAC 0.1143

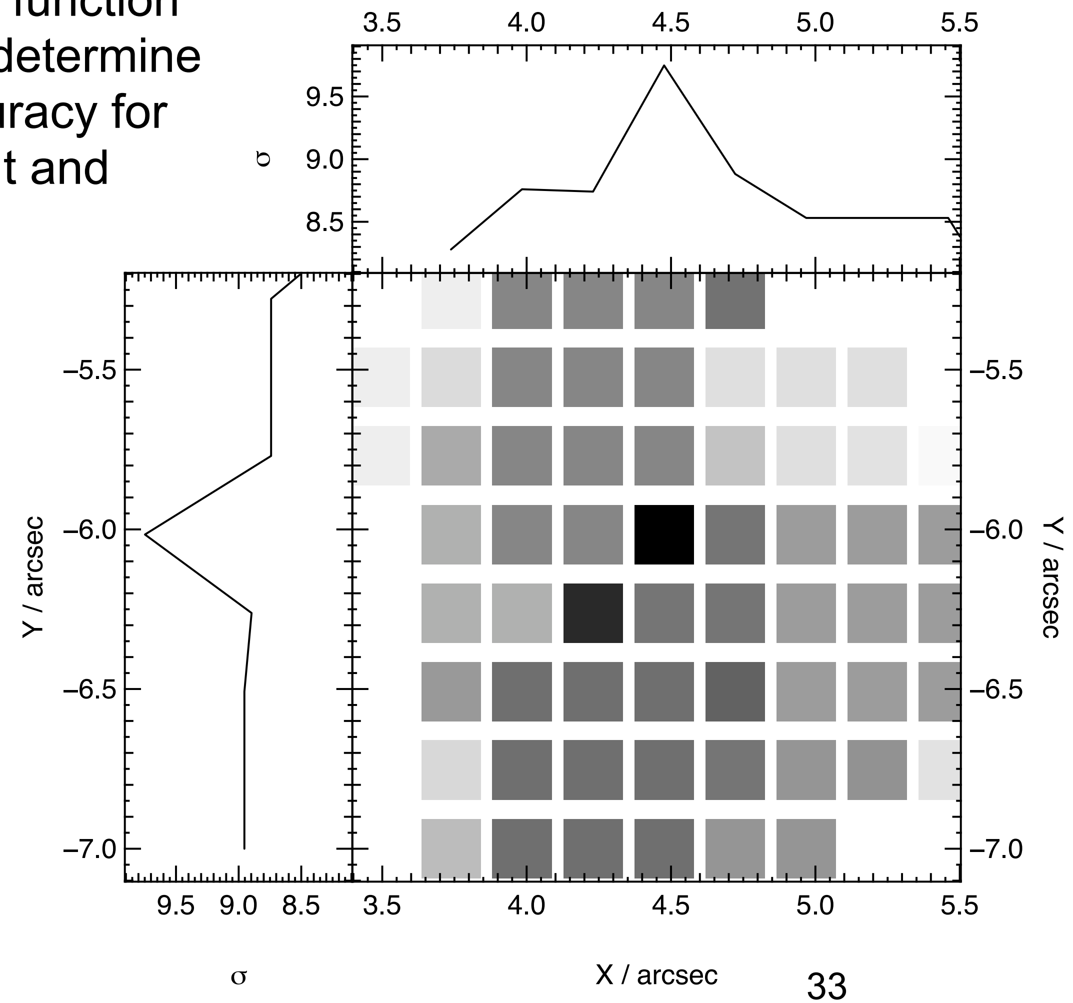

Dither 1
Dither 2
Dither 3
Row 5
Row 6

HET - Dec 08


SUM - 4465A, Sig=5.345, Chi2=1.188, Chi2s=1.701, CL/CR=221.5/356.8, F=740.4
Spesvphet0837.fits, FIBER 99, CCD 406.9,1223, FRAC 0.1132
Spesvphet0837.fits, FIBER 98, CCD 406.8,1231, FRAC 0.1112
Spesvphet0836.fits, FIBER 98, CCD 406.8,1231, FRAC 0.2016
Spesvphet0835.fits, FIBER 98, CCD 406.8,1231, FRAC 0.105

Determining Object Positions

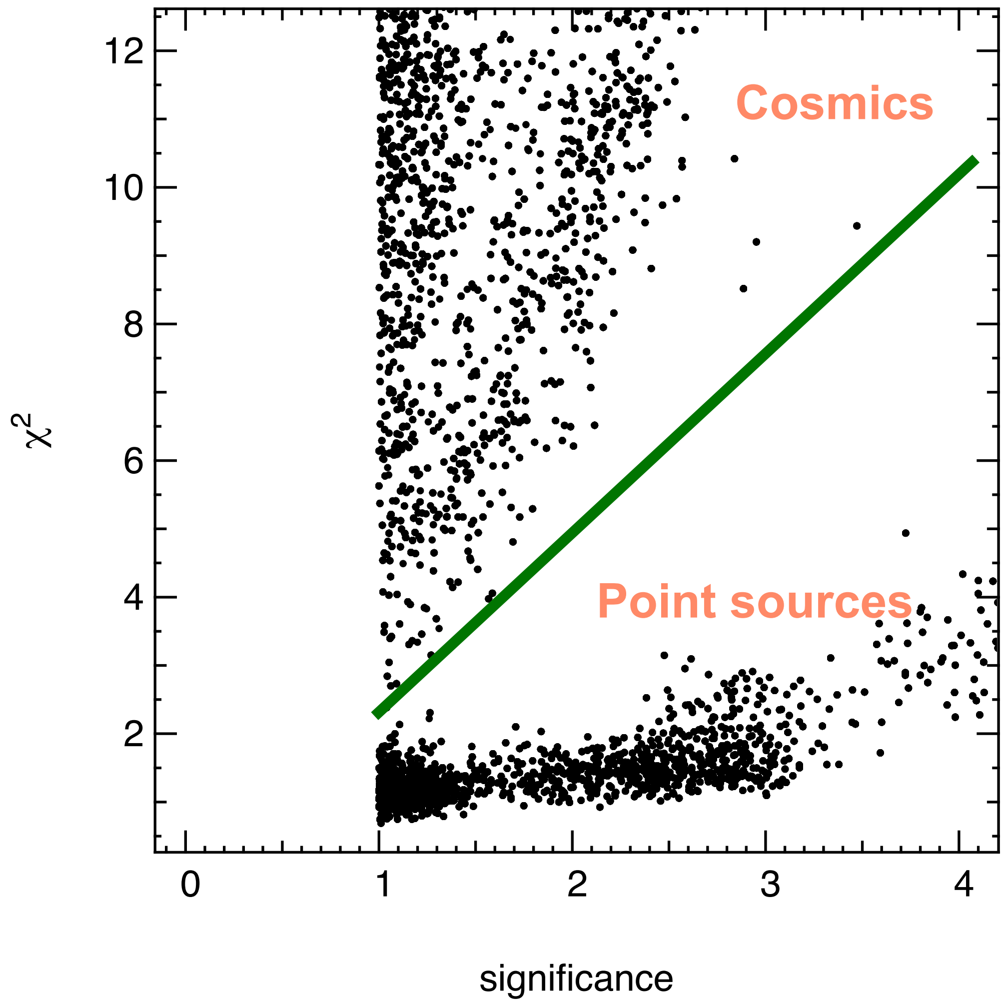
Distribution of σ and χ^2 as a function of sky {α,δ} can be used to determine source positions to ~1” accuracy for sources at the detection limit and ~0.4” for brighter sources.



Cosmic Rays

Cosmic rays have similar σ (fluxes) but very different χ^2.

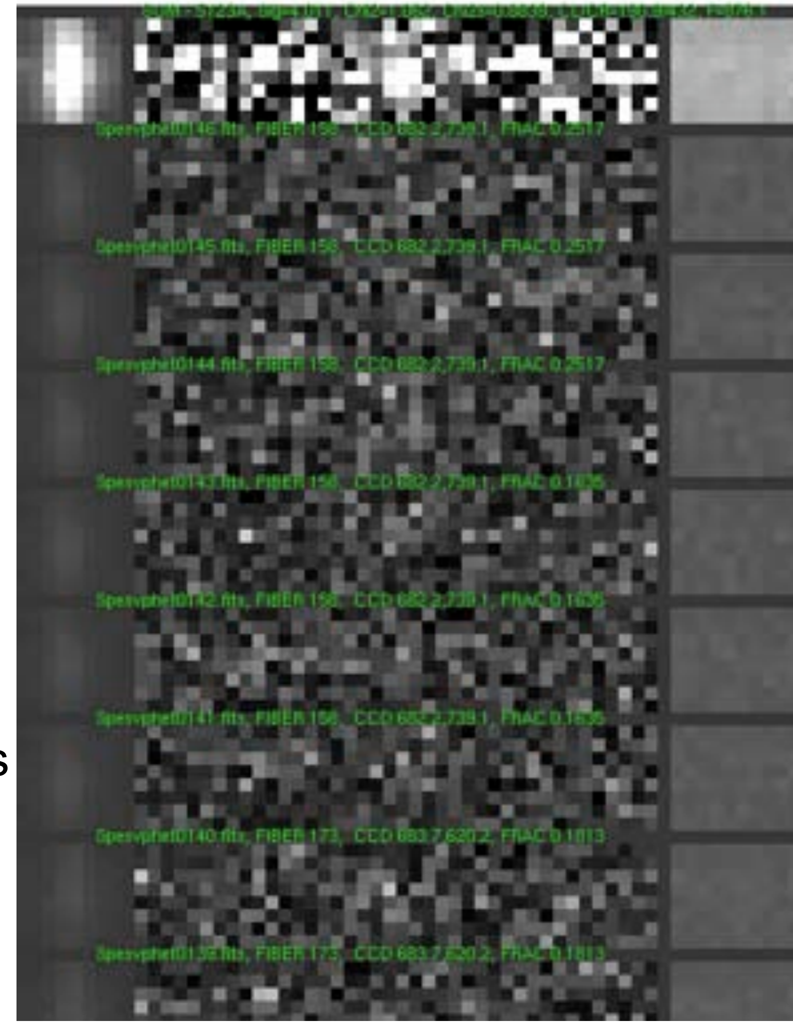
The plot shows simulations of LAEs added to real raw data and reduced with this pipeline.



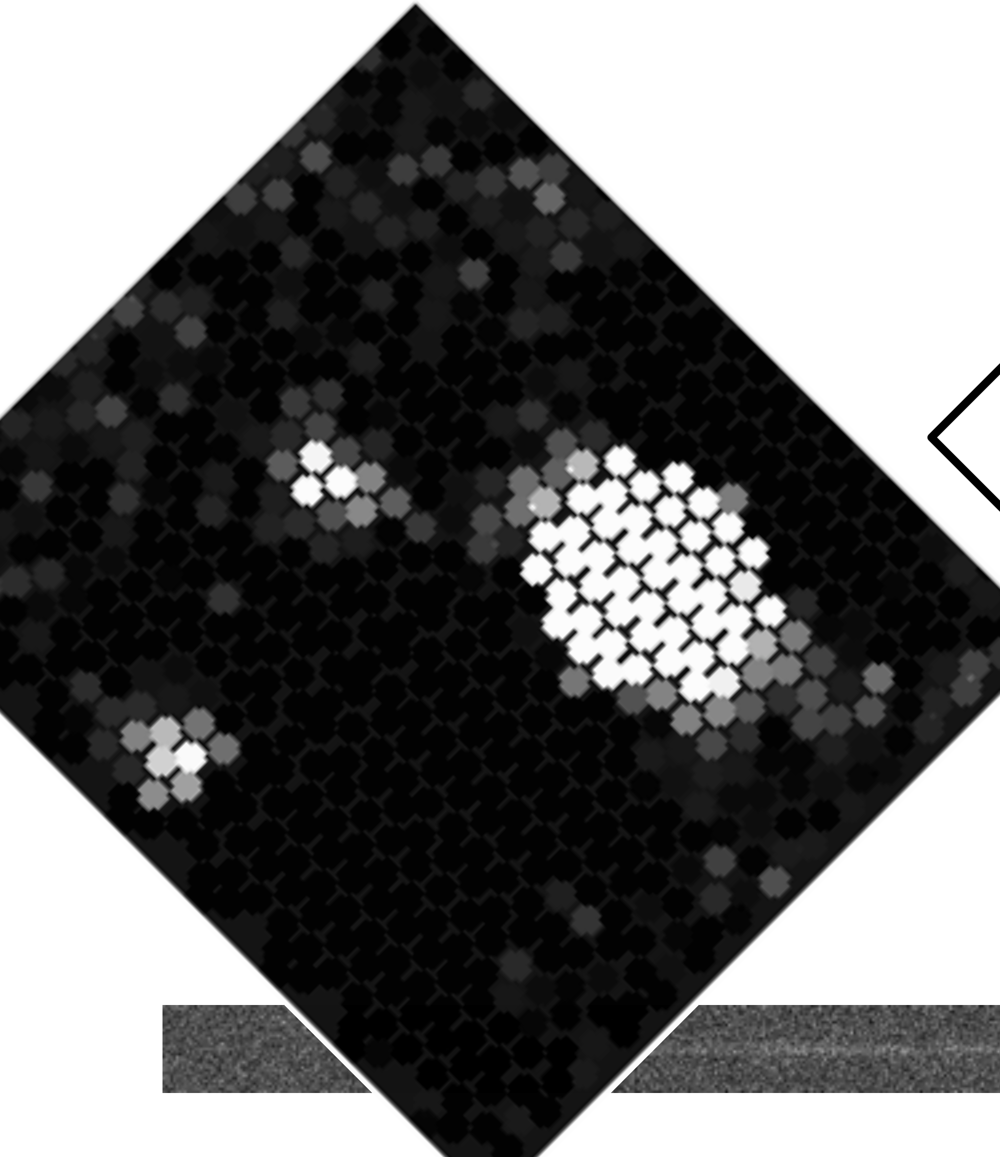
Detection Limits

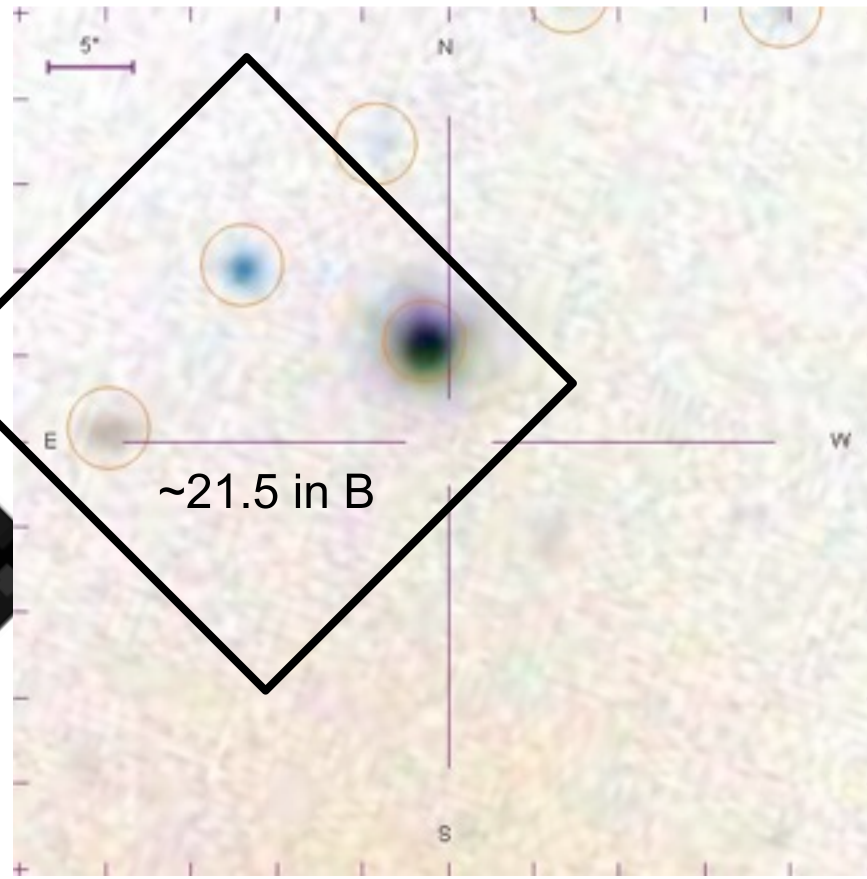
Define limit in analogous way to sensitivity simulations: total signal divided by total noise within resolution elements in N fibers.

- At 5-sigma, 0 cosmics are detected.
- Surprisingly, as one lowers sigma, the limiting factor becomes faint continuum sources.
- Cosmics still robustly excluded at 3.5 sigma.
- Full detection efficiency simulations are currently being performed by adding sources to the raw VP-HET data and reducing this data with the pipeline.
- Possible since the logic of the detection algorithm can be used to put objects into the data very easily.



Continuum Sources





Data Analysis Plan: Data Volume

HETDEX Data volume:

- 192 CCDs, 2 Mpix, 6 frames/shot = 9.2 GB/shot.
- 4000 shot survey, 4.6×10^6 science images.
- 36.8 TB raw science data + calibrations ~100TB total data volume.
- Include processing, multiple reductions: ~500TB storage.
- 2.76×10^4 science images/night = 221 GB/night science data volume.

- Quicklook: 392Mpix; 100Mpix/Min/CPU including database+I/O overhead; can be done on ~16 CPUs (today's performance).

- Full reduction: detection ~ 1Mpix/Min/CPU: 56GPix/night. To be reduced in 8h, need ~128 CPUs.

- compare PanSTARRS: 2TB/night; 7PB data volume, 300MB/s data rate sustained over 3 years ...

Non-DEX Requirements and Wish List

- The HETDEX pipeline is fairly specialized for DEX.
- Which non-line sources do we extract and how?
- AGN classification?
- Radial velocities for stars?
- How do we include user code?

- What facilities do we provide for non-DEX observations?
- How do we deal with large galaxies?
- Parallel mode? Who writes the software for this?
- non-DEX spectrograph configurations?
- Target selection for follow-up?